

Random modification effect in the size of the fluctuation of the LCS of two sequences of i.i.d. blocks

Heinrich Matzinger* Felipe Torres†

November 15, 2010

Abstract

The problem of the order of the fluctuation of the Longest Common Subsequence (LCS) of two independent sequences has been open for decades. There exist contradicting conjectures on the topic, [1] and [2]. In the present article, we consider a special model of i.i.d. sequences made out of blocks. A *block* is a contiguous substring consisting only of one type of symbol. Our model allows only three possible block lengths, each been equiprobable picked up. For i.i.d. sequences with equiprobable symbols, the blocks are independent of each other. For this model, we introduce a random operation (random modification) on the blocks of one of the sequences. In this article, for our block model, we show the techniques to prove the following: if we suppose that the random modification increases the length of the LCS with high probability, then the order of the fluctuation of the LCS is as conjectured by Waterman [2]. This result is a key technical part in the study of the size of the fluctuation of the LCS for sequences of i.i.d. blocks, developed in [3].

1 Model and main results

In general through this paper, X and Y will denote two finite strings over a finite alphabet Σ . A common subsequence of X and Y is a subsequence which is a subsequence of X as well as of Y . A Longest Common Subsequence of X and Y (denoted simply by LCS of X and Y , or only LCS when the context is clear enough) is a common subsequence of X and Y of maximal length. For a motivation on why to study the LCS problem, the reader can look at [3, 4].

Let $l > 0$ be an integer parameter. Let B_{X1}, B_{X2}, \dots and B_{Y1}, B_{Y2}, \dots be two i.i.d. sequences independent of each other such that:

$$\begin{aligned} P(B_{Xi} = l - 1) = P(B_{Xi} = l) = P(B_{Xi} = l + 1) &= 1/3 \\ P(B_{Yi} = l - 1) = P(B_{Yi} = l) = P(B_{Yi} = l + 1) &= 1/3. \end{aligned}$$

*School of Mathematics, Georgia Institute of Technology, 686 Cherry Street, GA 30332-0160 Atlanta, USA

†Fakultät für Mathematik, Universität Bielefeld, Postfach 100131 D-33501 Bielefeld, Germany

We call the runs of 0's and 1's blocks. Let $X^\infty = X_1X_2X_3\ldots$ be the binary sequence so that the i -th block has length B_{X_i} where X_1 is chosen 0 with probability 1/2 or 1 with probability 1/2. Similarly let $Y^\infty = Y_1Y_2Y_3\ldots$ be the binary sequence so that the i -th block has length B_{Y_i} and Y_1 is chosen 0 with probability 1/2 or 1 with probability 1/2.

Example 1.1 Assume that $X_1 = 0$ and $B_{X_1} = 3$, $B_{X_2} = 4$ and $B_{X_3} = 2$. Then we have that the sequence X^∞ starts as follows $X^\infty = 000111100\ldots$ meaning that in X^∞ the first block consists of three 0's, the second block consists of four 1's, the third block consists of two 0's, etc.

Let X denote the sequence obtained by only taking the first n bits of X^∞ , namely $X = X_1X_2X_3\ldots X_n$ and similarly $Y = Y_1Y_2Y_3\ldots Y_n$. Let L_n denote the length of the LCS of X and Y , $L_n := |\text{LCS}(X, Y)|$.

The main result of [3, 4] states that for l large enough, the order of the fluctuation of L_n is n :

Theorem 1.1 *There exists l_0 so that for all $l \geq l_0$ we have that:*

$$\text{VAR}[L_n] = \Theta(n)$$

for n large enough.

In [3, 4] the authors showed that theorem 1.1 is equivalent to proving that “a certain random modification has a biased effect on L_n ”. This is a technique with similar approaches in other papers (for instance see [5], [6]). So the main difficulty is actually proving that the random modification has typically a biased effect on the LCS, which for the block model is connected to a constrained optimization problem [3, 4]. This random modification is performed as follows: we choose at random in X a block of length $l - 1$ and at random one block of length $l + 1$, this means that all the blocks in X of length $l - 1$ have the same probability to be chosen and then we pick one of those blocks of length $l - 1$ up and also that all the blocks in X of length $l + 1$ have the same probability to be chosen and we pick one of those blocks of length $l + 1$ up. Then we change the length of both these blocks to l . The resulting new sequence is denoted by \tilde{X} . Let \tilde{L}_n denote the length of the LCS after our modification of X . Hence:

$$\tilde{L}_n := |\text{LCS}(\tilde{X}, Y)|.$$

If we can prove that our block length changing operation has typically a biased effect on the LCS than the order of the fluctuation of L_n is \sqrt{n} . This is the content of the next theorem:

Theorem 1.2 *Assume that there exists $\epsilon > 0$ and $\alpha > 0$ not depending on n such that for all n large enough we have:*

$$\mathbb{P} \left(\mathbb{E}[\tilde{L}_n - L_n | X, Y] \geq \epsilon \right) \geq 1 - \exp(-n^\alpha). \quad (1.1)$$

Then,

$$\text{VAR}[L_n] = \Theta(n)$$

for n large enough.

The above theorem reduces the problem of the order of fluctuation to proving that our random modification has typically a higher probability to lead to an increase than to a decrease in score. **The main result of this article is theorem 1.2.**

A very useful tool we often use is the Azuma-Hoeffding theorem. The following is a version of it for martingales (for a proof see [7]):

Theorem 1.3 (*Hoeffding's inequality*) *Let (V, \mathfrak{F}) be a martingale, and suppose that there exists a sequence $\mathfrak{a}_1, \mathfrak{a}_2, \dots$ of real numbers such that*

$$\mathbb{P}(|V_n - V_{n-1}| \leq \mathfrak{a}_n) = 1$$

for all n . Then:

$$\mathbb{P}(|V_n - V_0| \geq v) \leq 2 \exp \left\{ -\frac{1}{2} v^2 / \sum_{i=1}^n \mathfrak{a}_i^2 \right\} \quad (1.2)$$

for every $v > 0$.

We also will use a corollary of the above theorem, for some intermediate bounds:

Corollary 1.1 *Let $a > 0$ be constant and V_1, V_2, \dots be an i.i.d sequence of random bounded variables such that:*

$$\mathbb{P}(|V_i - \mathbb{E}[V_i]| \leq a) = 1$$

for every $i = 1, 2, \dots$. Then for every $\Delta > 0$, we have that:

$$\mathbb{P} \left(\left| \frac{V_1 + \dots + V_n}{n} - \mathbb{E}[V_1] \right| \geq \Delta \right) \leq 2 \exp \left(-\frac{\Delta^2}{2a^2} \cdot n \right) \quad (1.3)$$

2 Random modification effect in the fluctuation

We are going to prove theorem 1.2 which states that $\text{VAR}[L_n] = \Theta(n)$ holds if there exist $\epsilon, \alpha > 0$ not depending on n such that:

$$\mathbb{P} \left(\mathbb{E}[\tilde{L}_n - L_n | X, Y] \geq \epsilon \right) \geq 1 - \exp(-n^\alpha). \quad (2.1)$$

for all n large enough. We have omitted some of the proofs for shortness reasons, but all the details can be looked at [4].

Note that if \mathcal{Z} is a random variable with $\text{VAR}[\mathcal{Z}] = \Theta(n)$ and f is a map which tends to increase linearly, then for $\mathcal{W} = f(\mathcal{Z})$, we also have the order $\text{VAR}[\mathcal{W}] = \Theta(n)$. The map f can be even a random map but must be independent of \mathcal{Z} . The exact basic result ([6], lemma 3.2) goes as follows:

Lemma 2.1 *Let $c > 0$ be a constant. Assume that $g : \mathbb{R} \rightarrow \mathbb{R}$ is a map which is everywhere differentiable and such that for all $x \in \mathbb{R}$ we have:*

$$\frac{dg(x)}{dx} \geq c.$$

Let B be a random variable such that $E[|g(B)|] < +\infty$. Then:

$$\text{VAR}[g(B)] \geq c^2 \cdot \text{VAR}[B].$$

In the present context, we need a slightly different version:

Lemma 2.2 *Let $\epsilon, m > 0$ be constants and $f : \mathbb{Z} \rightarrow \mathbb{Z}$ be a map such that for all $z_1 \leq z_2$ the following two conditions hold:*

$$z_2 - z_1 \geq m \Rightarrow f(z_2) - f(z_1) \geq \frac{\epsilon}{8}(z_2 - z_1) \quad (2.2)$$

$$\exists \beta > 0 : z_2 - z_1 < m \Rightarrow f(z_2) - f(z_1) \leq \beta(z_2 - z_1) \quad (2.3)$$

Let B be a random variable such that $E[|f(B)|] \leq +\infty$. Then:

$$\text{VAR}[f(B)] \geq \frac{\epsilon^2}{64} \left(1 - 16 \frac{(\epsilon/8 + \beta)m}{\epsilon \sqrt{\text{VAR}[B]}} \right) \text{VAR}[B] \quad (2.4)$$

Proof. Let $h : \mathbb{Z} \rightarrow \mathbb{Z}$ be a map defined from f as follows: for a given $z \in \mathbb{Z}$ choose $k \geq 2$ such that $z \in [km, (k+1)m]$ and compute

$$h(z) = \left(\frac{f((k+1)m) - f(km)}{m} \right) (z - km) + f(km)$$

then $h(z)$ is just the linear interpolation of $f(z)$ in $[km, (k+1)m]$. It is easy to see that h satisfies the conditions of lemma 2.1 for $c = \epsilon/8$. Then:

$$\text{VAR}[h(B)] \geq \frac{\epsilon^2}{64} \text{VAR}[B] \quad (2.5)$$

We want to estimate the distance between the random variables $h(B)$ and $f(B)$. First, we note that from 2.2 and by the definition of h , the following inequalities hold for $km \leq B \leq (k+1)m$:

$$\frac{\epsilon}{8}(B - km) + f(km) \leq f(B), h(B) \leq \frac{\epsilon}{8}(B - (k+1)m) + f((k+1)m)$$

looking at conditions 2.2, 2.3 and the inequalities above we get

$$\begin{aligned} |h(B) - f(B)| &\leq \left| \frac{\epsilon}{8}(B - km) + f(km) - \frac{\epsilon}{8}(B - (k+1)m) + f((k+1)m) \right| \\ &\leq \frac{\epsilon}{8}m + |f((k+1)m) - f(km)| \\ &\leq \left(\frac{\epsilon}{8} + \beta \right) m \end{aligned}$$

and by using the last inequality above:

$$\text{VAR}[f(B) - h(B)] \leq \left(\frac{\epsilon}{8} + \beta \right)^2 m^2. \quad (2.6)$$

Since $f(B) = h(B) + (f(B) - h(B))$ we can apply triangular inequality and find:

$$\sqrt{\text{VAR}[f(B)]} \geq \sqrt{\text{VAR}[h(B)]} - \sqrt{\text{VAR}[f(B) - h(B)]},$$

hence we have:

$$\begin{aligned} \text{VAR}[(f(B))] &\geq \text{VAR}[h(B)] - 2\sqrt{\text{VAR}[h(B)]} \cdot \sqrt{\text{VAR}[f(B) - h(B)]} \\ &= \text{VAR}[h(B)] \left(1 - 2 \frac{\sqrt{\text{VAR}[f(B) - h(B)]}}{\sqrt{\text{VAR}[h(B)]}} \right) \end{aligned}$$

Finally, applying the inequalities 2.5 and 2.6 to the last inequality above, we get:

$$\text{VAR}[(f(B))] \geq \frac{\epsilon^2}{64} \left(1 - 16 \frac{(\epsilon/8 + \beta)m}{\epsilon \sqrt{\text{VAR}[B]}} \right) \text{VAR}[B]. \quad \blacksquare$$

Hence to prove that $\text{VAR}[L_n] = \Theta(n)$, we try to represent L_n as $f(\mathcal{Z})$ where f is a random map which tends to increase linearly on a certain scale and \mathcal{Z} is a random variable having fluctuation of order \sqrt{n} .

2.1 Random modifications and the variables (T, Z, R)

Let N_l denote the number of blocks in X of length l , whilst N_{l-1} , resp. N_{l+1} denote the number of blocks of length $l-1$, resp $l+1$ in X . Let us define the following three random variables:

$$T := N_l + N_{l-1} + N_{l+1} \tag{2.7}$$

$$Z := N_l - N_{l-1} - N_{l+1} \tag{2.8}$$

$$R := n - (l N_l + (l+1) N_{l+1} + (l-1) N_{l-1}) \tag{2.9}$$

Note that when we know the values of (T, Z, R) we can determine the values of N_{l-1}, N_l and N_{l+1} as a linear function by using the definitions of T, Z and R as follows:

$$\begin{pmatrix} N_{l-1}(T, Z, R) \\ N_l(T, Z, R) \\ N_{l+1}(T, Z, R) \end{pmatrix} = \begin{pmatrix} (2l+1)/4 & -1/4 \\ 1/2 & 1/2 \\ -(2l-1)/4 & -1/4 \end{pmatrix} \begin{pmatrix} T \\ Z \end{pmatrix} + \begin{pmatrix} -(n-R)/2 \\ 0 \\ (n-R)/2 \end{pmatrix} \tag{2.10}$$

The variable R represents what is left in X after the last block of length $l-1, l$ or $l+1$.

Example 2.1 *Let us consider the sequence $X = 000111100011001$ for $l = 3$ and $n = 15$. We see that $N_{l-1} = 2$, $N_l = 2$ and $N_{l+1} = 1$, hence $T = 5, Z = -1$ and $R = 1$. Also, the block 1 at the end of X has length strictly smaller than $l-1$ which also means that $R = 1$. In this case is easy to interpret what R is since the last block in X has length strictly less than $l-1$. Let us see a different situation. Let us take again $l = 3$ and now consider $B_{X1} = 2, B_{X2} = 3, B_{X3} = 4, B_{X4} = 3, B_{X5} = 2, B_{X6} = 4, \dots$ such that $X^\infty = 001110000111001111 \dots$ Take $n = 16$ so that $X = 0011100001110011$. Here the last block of X has length $l-1 = 2$ which should imply (using the point of view of the last situation) that $R = 0$. But, notice that the*

block in X^∞ corresponding to B_{X_6} was cut when we took X . In this case, we say that the last block in X corresponds to the rest so $R = 2$ and therefore $N_{l-1} = 2, N_l = 2$ and $N_{l+1} = 1$, then $T = 5$ and $Z = -1$. We take this convention on R , even if the definition 2.9 is not the exact one, because of the simplifications later during the computation of the joint distribution of N_{l-1}, N_l, N_{l+1} .

Let us roughly explain the main idea behind this subsection. Assume that we have a random couple (V, W) which can take on a finite number of values only. We also assume the joint distribution $\mathcal{L}(V, W)$ to be given. To simulate (V, W) , we could first simulate V using the marginal law $\mathcal{L}(V)$. We would obtain a numeric value v_0 . Then, we could simulate W using the conditional law $\mathcal{L}(W|V = v_0)$ and obtain the numeric value w_0 . The couple (v_0, w_0) has joint distribution $\mathcal{L}(V, W)$. Another less efficient possibility is to simulate for each (non-random) value v that V can take, a value for W with distribution $\mathcal{L}(W|V = v)$. Call the numeric value $w(v)$. Then, we would simulate V with distribution $\mathcal{L}(V)$ and obtain a numeric value v_0 . Then, for W we would take among all the values which we have simulated, the one corresponding to $V = v_0$. In this manner, we get $(v_0, w(v_0))$. This couple has the distribution $\mathcal{L}(V, W)$ and this does not even require that we simulate the different $w(v)$'s independently of each other. Only, V needs to be simulated independently of the assignment $v \mapsto w(v)$.

We are going to do the above simulation scheme with V being (T, Z, R) and W being the rest of the information in (X, Y) . More precisely, for all possible (t, z, r) non-random values, we simulate X conditional on $(T, Z, R) = (t, z, r)$. The resulting string is denoted by $X_{(t,z,r)}$ and has thus distribution

$$\mathcal{L}(X_{(t,z,r)}) = \mathcal{L}(X \mid (T, Z, R) = (t, z, r)).$$

Let $L_n(t, z, r)$ denote the length of the LCS

$$L_n(t, z, r) := |\text{LCS}(X_{(t,z,r)}, Y)|.$$

We assume that the simulation of the string $X_{(t,z,r)}$ is done independently of (T, R, Z) and of Y . In this manner, we get that $L_n(T, Z, R)$ has same distribution as $L_n = |\text{LCS}(X, Y)|$. So to prove that $\text{VAR}[L_n] = \Theta(n)$, it is enough to prove that

$$\text{VAR}[L_n(T, Z, R)] = \Theta(n). \tag{2.11}$$

We saw at the beginning of this section (see lemma 2.1 and 2.2), that when we transform a variable having variance of order $\Theta(n)$ with a map which tends to increase linearly, then the resulting variable has variance of order $\Theta(n)$. It is easy to see that $\text{VAR}[Z] = \Theta(n)$ (see also lemma 2.8). Hence to prove 2.11, it is enough to show that with high probability the (random) map

$$z \mapsto L_n(T, z, R)$$

tends to increase linearly (on the appropriate scale and on a domain on which Z typically takes its value). That means, we need to show that we can simulate the values $L_n(t, z, r)$ in such a manner to get the desired distribution $\mathcal{L}(X|(T, Z, R) = (t, z, r))$ as well as the desired linear increase of the map $z \mapsto L_n(T, z, R)$. This is achieved by simulating $X_{(t,z,r)}$ in the

following way: for a given value (t, r) , so that $P((T, R) = (t, r)) \neq 0$, we take a left most (left most to be defined later) value z_0 and simulate a string with distribution equal to the conditional distribution of X given $(T, Z, R) = (t, z_0, r)$. That resulting string is denoted by $X_{(t, z_0, r)}$. Then, we apply the random modification to $X_{(t, z_0, r)}$. This means, we choose one block of length $l - 1$ and one block of length $l + 1$ at random in $X_{(t, z_0, r)}$ and turn them both into length l . The resulting string is denoted by $X_{(t, z_0+4, r)}$. Then, we choose at random in $X_{(t, z_0+4, r)}$ a block of length $l - 1$ and a block of length $l + 1$ and turn them both into length l . The new string which we obtain in this manner is denoted by $X_{(t, z_0+8, r)}$. We keep repeating this same operation to obtain the sequence of strings

$$X_{(t, z_0, r)}, X_{(t, z_0+4, r)}, X_{(t, z_0+8, r)}, \dots \quad (2.12)$$

For each value of (t, r) with $P((T, R) = (t, r)) \neq 0$ we obtain two finite sequences of strings: first 2.12 and then

$$X_{(t, z_0+2, r)}, X_{(t, z_0+6, r)}, X_{(t, z_0+10, r)}, \dots$$

by a similar procedure. Namely, after $X_{(t, z_0+2, r)}$ is generated with distribution X conditional on $(T, Z, R) = (t, z + 2, R)$, the subsequent strings are obtained by applying sucessively the random modification tilde, which chooses at random in the string a block of length $l - 1$ and a block of length $l + 1$ and turn them both into length l .

Recall that in this section we assume that our random modification has a biased effect of $\epsilon > 0$ on the LCS, so that with high probability

$$E[\tilde{L}_n - L_n \mid X, Y] \geq \epsilon.$$

Hence, it follows that the map $z \mapsto L_n(T, z, R)$ tends with high probability to increase with slope close to ϵ on a constant time scale $\ln n$ (the constant must be taken large enough though, see lemma 2.6 and proposition 2.2). In other words, since the random modification has a bi-ased positive effect, the map $z \mapsto L_n(T, z, R)$ behaves like a random walk with drift ϵ . The only thing which remains to be proved is that with our scheme of using the random modification, the strings $X_{(t, z, r)}$ have the right distribution, i.e. the distribution of X conditional on $(T, Z, R) = (t, z, r)$. This is proved in lemma 2.5.

We have so far summarized the idea which explains why the biased effect of the random modification implies $\text{VAR}[L_n] = \Theta(n)$. There is one more detail which we should mention and which makes notations a little more difficult. To prove that $z \mapsto L_n(T, z, R)$ tends to increase linearly we use the biased effect on the LCS for the random modification. However, this bias holds with high probability for X and not for $X_{(t, z, r)}$. When we look at the conditional distribution of X given $(T, Z, R) = (t, z, r)$, we divide by the probability

$$P((T, Z, R) = (t, z, r)). \quad (2.13)$$

The string $X_{(t, z, r)}$ has distribution of X conditional on $(T, Z, R) = (t, z, r)$. So for the biased effect to have large probability also for $X_{(t, z, r)}$ (and not just for X), we need the probability 2.13 to not be too small. To assure this, we will restrict ourselves to “typical” values for (T, Z, R) . We will consider only values for (T, Z) which lie in an interval $D = D_T \times D_Z$ (see

definition below 2.16) and prove that any possible value $(t, z) \in D_z \times D_t$ has polynomially bounded probability (see lemma 2.4).

Let us now give all the details:

Proposition 2.1 *Given $\epsilon > 0$ there exist constants $1 \leq k_1, k_2, k_3 \leq k^*$ all not depending on n but on ϵ such that:*

$$\mathbb{P} \left(\left| \frac{N_{l-1} - \frac{n}{3l}}{\sqrt{n}} \right| \leq k_1 \right), \quad \mathbb{P} \left(\left| \frac{N_l - \frac{n}{3l}}{\sqrt{n}} \right| \leq k_2 \right), \quad \mathbb{P} \left(\left| \frac{N_{l+1} - \frac{n}{3l}}{\sqrt{n}} \right| \leq k_3 \right) \geq 1 - \epsilon \quad (2.14)$$

for every n large enough.

We will need later the following lemma:

Lemma 2.3 *There exists $c > 0$ not depending on n such that:*

$$\mathbb{P} \left(T \in \left[\frac{n}{l} - c\sqrt{n}, \frac{n}{l} + c\sqrt{n} \right], Z \in \left[-\frac{n}{3l} - c\sqrt{n}, -\frac{n}{3l} + c\sqrt{n} \right] \right) \geq 0.9 \quad (2.15)$$

Let D denote the domain

$$D := \left[\frac{n}{l} - c\sqrt{n}, \frac{n}{l} + c\sqrt{n} \right] \times \left[-\frac{n}{3l} - c\sqrt{n}, -\frac{n}{3l} + c\sqrt{n} \right] \quad (2.16)$$

and let

$$\begin{aligned} D_T &:= \left[\frac{n}{l} - c\sqrt{n}, \frac{n}{l} + c\sqrt{n} \right] \\ D_Z &:= \left[-\frac{n}{3l} - c\sqrt{n}, -\frac{n}{3l} + c\sqrt{n} \right] \end{aligned}$$

hence,

$$D = D_T \times D_Z.$$

Given $(t, z) \in D$ such that $(T, Z) = (t, z)$ we have

$$N_{l-1}(t, z) + N_l(t, z) + N_{l+1}(t, z) = t.$$

The probability for a realization of N_{l-1}, N_l and N_{l+1} is given by:

$$\begin{aligned} \mathbb{P}(T = t, Z = z, R = r) &= \binom{N_{l-1}(t, z) + N_l(t, z) + N_{l+1}(t, z)}{N_{l-1}(t, z) \quad N_l(t, z) \quad N_{l+1}(t, z)} \left(\frac{1}{3} \right)^t \cdot \mathbb{P}(B_{X1} > r) \\ &= \frac{t!}{(N_{l-1}(t, z))! (N_l(t, z))! (N_{l+1}(t, z))!} \left(\frac{1}{3} \right)^t \cdot \mathbb{P}(B_{X1} > r) \end{aligned} \quad (2.17)$$

where the probability $\mathbb{P}(B_{X1} > r) = \mathbb{P}(R = r)$ is due to the convention of R described in the example 2.1. Finally, due to 2.10, for any $n_1, n_2, n_3 \in \mathbb{N}$ the conditional joint distribution

$$\mathbb{P}(N_{l-1}(T, Z, R) = n_1, N_l(T, Z, R) = n_2, N_{l+1}(T, Z, R) = n_3 \mid R = r)$$

is multinomial.

Lemma 2.4 *There exists $k_0 > 0$ not depending on n (but depending on c) such that for every $(t, z) \in D$ and $r < l + 1$ for which the probability $P((T, Z, R) = (t, z, r)) \neq 0$, we have that:*

$$P((T, Z, R) = (t, z, r)) \geq \frac{k_0}{n}$$

for every n large enough.

Note that for any variables X and Y we have (see for example [8])

$$\text{VAR}[Y] = E[\text{VAR}[Y|X]] + \text{VAR}[E[Y|X]] \geq E[\text{VAR}[Y|X]]. \quad (2.18)$$

Let O be the random variable which is equal to one when (T, Z) is in D and 0 otherwise. We can now use inequalities 2.15 and 2.18 to find

$$\text{VAR}[L_n] \geq E[\text{VAR}[L_n|O]] \geq \text{VAR}[L_n|O = 1] \cdot P(O = 1) \geq 0.9\text{VAR}[L_n|O = 1] \quad (2.19)$$

Next for every (t, z) in D and $r < l + 1$ we are going to simulate the random variable L_n conditional on $(T, Z, R) = (t, z, r)$. We denote the result by $L_n(t, z, r)$. In other words, the distribution of $L_n(t, z)$ is equal to

$$\mathcal{L}(L_n(t, z, r)) = \mathcal{L}(L_n|(T, Z, R) = (t, z, r)).$$

Let (T_D, Z_D) denote a variable having the distribution of (T, Z) conditional on the event $(T, Z) \in D$. We assume that all the $L_n(t, z, r)$ are independent of (T_D, Z_D) . Then, we get that

$$L_n(T_D, Z_D, R)$$

has same distribution as L_n conditional on $(T, Z) \in D$. Hence, we get

$$\text{VAR}[L_n|O = 1] = \text{VAR}[L_n(T_D, Z_D, R)] \quad (2.20)$$

By using 2.18, we find

$$\text{VAR}[L_n(T_D, Z_D, R)] \geq E[\text{VAR}[L_n(T_D, Z_D, R)|T_D, R]]. \quad (2.21)$$

Note that for $L_n(T_D, Z_D, R)$ to have the same distribution as L_n conditional on $(T, Z) \in D$ and on $R = r$, the variables $L_n(t, z, r)$ do not need to be independent of each other. We are next going to explain how we simulate the variables $L_n(t, z, r)$ a bit more in detail as before. We simulate a string $X_{(t, z, r)}$ having the distribution of the string X conditional on the event $(T, Z, R) = (t, z, r)$. Then we put

$$L_n(t, z, r) = |\text{LCS}(X_{(t, z, r)}, Y)|.$$

Next, let us describe how we simulate $X_{(t, z, r)}$ based on what was roughly explained at the beginning of subsection 2.1. Given $t_0 \in D_T$ the most left element in D_T and $r_0 < l - 1$, we are going to simulate $X_{(t_0, z, r_0)}$ for $z \in D_Z$ only if $P((T, Z, R) = (t_0, z, r_0)) \neq 0$. We simulate $X_{(t_0, z_0, r_0)}$ so that it has distribution $\mathcal{L}(X|(T, Z, R) = (t_0, z_0, r_0))$. Next, we simulate $X_{(t_0, z_0+2, r_0)}$ by choosing in X , with the same probability, a block of length $l - 1$ either a block

of length $l + 1$ and change its length to l . The next realization we simulate is $X_{(t_0, z_0+4, r_0)}$ by choosing in X , with the same probability, a block of length $l - 1$ and a block of length $l + 1$ and change their lengths to l (this is our usual random modification). Then by induction we simulate

$$\{X_{(t_0, z_0+4i, r_0)} : i = 1, 2, \dots\}$$

with our usual random modification and later

$$\{X_{(t_0, z_0+2+4i, r_0)} : i = 1, 2, \dots\}$$

just starting with $X_{(t_0, z_0+2, r_0)}$ and performing our usual random modification to get each $X_{(t_0, z_0+6, r_0)}, X_{(t_0, z_0+10, r_0)}, X_{(t_0, z_0+14, r_0)}$, etc. Both inductions run until indexes i_0 , resp. i_0^* , satisfying:

$$\begin{aligned} z_0 + 4i_0 &\leq -\frac{n}{3l} + c\sqrt{n} \Rightarrow i_0 \leq \sqrt{n} \\ z_0 + 2 + 4i_0^* &\leq -\frac{n}{3l} + c\sqrt{n} \Rightarrow i_0^* \leq \frac{\sqrt{n} - 1}{2} \end{aligned}$$

For simplicity, let us call $z_0, z_1 = z_0 + 2, z_2 = z_0 + 4, \dots, z_d$ all the values which Z takes. After we have simulated $X_{(t_0, z_0, r_0)}, X_{(t_0, z_1, r_0)}, \dots, X_{(t_0, z_d, r_0)}$ we fix $t_1 = t_0 + 1$ and repeat all the procedure again starting with the simulation of $X_{(t_1, z_0, r_0)}$. We keep taking $t_2 < t_3 < t_4 \dots$ all natural numbers in D_T to finish all the simulation of $\{X_{(t, z, r_0)} : t \in D_T, z = z_0, z_1, \dots, z_d\}$. Once we have finished with that, we take $r_1 < l - 1$ natural number and do all the simulation above starting with $X_{(t_0, z_0, r_1)}$ only if $P((t_0, z_0, r_1)) \neq 0$. Finally, we obtain the complete sequence $\{X_{(t, z, r)} : t \in D_T, z = z_0, z_1, \dots, z_d, r = 0, \dots, l - 2\}$, where each (t, z, r) has probability $P((T, Z, R) = (t, z, r)) \neq 0$.

We need to verify that this operation give us the equiprobable distribution. This is the content of the next lemma:

Lemma 2.5 *Assume that $X_{(t, z, r)}$ is distributed according to*

$$\mathcal{L}(X|(T, Z, R) = (t, z, r)).$$

Choose at random (with equal probability) in the string $X_{(t, z, r)}$ a block of length $l + 1$ and $l - 1$ and modify them to have both length l . Then the resulting string has distribution

$$\mathcal{L}(X|(T, Z, R) = (t, z + 4, r)).$$

Proof. Because of our linear equation system 2.10, we have that conditioning on T, Z, R is equivalent to conditioning on (N_{l-1}, N_l, N_{l+1}) . As mentioned, $X_{(t, z, r)}$ denotes a string of length n , having the distribution of X conditional on $(T, Z, R) = (t, z, r)$. We denote by $\tilde{X}_{(t, z, r)}$ the string we obtain by performing our random modification on $X_{(t, z, r)}$. In other words, $\tilde{X}_{(t, z, r)}$ is obtained by choosing a block of length $l + 1$ and a block of length $l - 1$ at random in $X_{(t, z, r)}$ and changing them both to length l . Let (n_1, n_2, n_3) be the number of blocks of length $l - 1, l$ and $l + 1$ corresponding to (t, z, r) . In other words, n_1, n_2 and n_3

are given by the linear system of equation 2.10 when $N_{l-1} = n_1, N_l = n_2, N_{l+1} = n_3$ and $T = t, Z = z, R = r$. We have

$$P(N_1 = n_1, N_2 = n_2, N_3 = n_3 | T = t, Z = z, R = r) = 1.$$

The distribution of the random string $X_{(t,z,r)}$ is the uniform distribution on $\xi^n(t, z, r)$. Here, $\xi^n(t, z, r)$ denotes the set of strings of length n , which consists only of blocks of length $l-1$, l and $l+1$, such that the total number of blocks is t , whilst the number of blocks of length l minus the number of blocks of length $l-1$ and $l+1$ is z . We also request that the rest block at the end has length r . We can describe $\xi^n(t, z, r)$ equivalently as the set of all strings consisting exactly of n_1 blocks of length $l-1$, n_2 blocks of length l and n_3 blocks of length $l+1$, no other blocks allowed except a rest block at the end which has length strictly less than $l-1$. In other words, the random string $X_{(t,z,r)}$ is such that the number of blocks of length $l-1$, l and $l+1$ is determined, only the order in which these blocks appear varies. Among others, each possible realization for $X_{(t,z,r)}$ which has non-zero probability has the same probability:

$$\binom{n_1 + n_2 + n_3}{n_1 \ n_2 \ n_3}^{-1} \quad (2.22)$$

When we apply the random modification, the variable T stays the same, the variable Z increases by 4 and the variable R stays the same.

Since the distribution of X conditional on (T, Z, R) is the uniform distribution on the appropriate set of strings, we have the following: for proving that $\tilde{X}_{(t,z,r)}$ has distribution of X conditional on $(T, Z, R) = (t, z+4, r)$ it is enough to show that its distribution is the uniform distribution on $\xi^n(t, z+4, r)$. For this, let \tilde{x} denote a (non-random) element of $\xi^n(t, z+4, r)$. Hence, the number of blocks in \tilde{x} of length $l-1$, l , resp $l+1$ is $n_1 - 1$, $n_2 + 2$, resp. $n_3 - 1$. The probability

$$P(\tilde{X}_{(t,z,r)} = \tilde{x})$$

can be calculated as follows: if we only know \tilde{x} , any block of length l of \tilde{x} could be the block which had length $l-1$ and has been turned into length l by the *tilde operation* (choosing blocks at random and changing their lengths). Same thing for the block which had length $l+1$. But when we know these two blocks, then the string before the random modification is uniquely determined. Let x be such a string which could lead to \tilde{x} after the random modification. There are hence $\tilde{n}_2 \cdot (\tilde{n}_2 - 1)$ such strings (here, $\tilde{n}_2 = n_2 + 2$, so that \tilde{n}_2 denotes the number of blocks of length l in \tilde{x}). The probability, given $X_{(t,z,r)} = x$, that the random string turns out to be \tilde{x} is equal to $1/(n_1 \cdot n_3)$. As a matter of fact, among the n_1 blocks of length $l-1$, there is exactly one which needs to be randomly modified. Similarly, among the n_3 blocks of length $l+1$, there is exactly one which needs to be changed into length l in order to obtain the string \tilde{x} . Hence,

$$P(\tilde{X}_{(t,z,r)} = \tilde{x} | X_{(t,z,r)} = x) = \frac{1}{n_1 \cdot n_3}. \quad (2.23)$$

Let ξ^{n*} denote the set of all strings which could lead to \tilde{x} if we apply the random modification to them. We saw that there are $(n_2 + 2)(n_2 + 1)$ elements in the set ξ^{n*} . By law of total

probability, we have

$$P(\tilde{X}_{(t,z,r)} = \tilde{x}) = \sum_{x \in \xi^{n*}} P(\tilde{X} = \tilde{x} | X = x) P(X_{(t,z,r)} = x) = \sum_{x \in \xi^{n*}} \frac{1}{n_1 \cdot n_3} \binom{n_1 + n_2 + n_3}{n_1 \ n_2 \ n_3}^{-1} \quad (2.24)$$

The last equation above was obtained using 2.23 and 2.22. Note that the sum on the most right of equation in 2.24, is a sum of $(n_2 + 2)(n_2 + 1)$ equal terms. This leads to

$$P(\tilde{X}_{(t,z,r)} = \tilde{x}) = \frac{(n_2 + 2)(n_2 + 1)}{n_1 \cdot n_3} \binom{n_1 + n_2 + n_3}{n_1 \ n_2 \ n_3}^{-1}.$$

The formula on the right side above does not depend on \tilde{x} . Hence, this proves that $\tilde{X}_{(t,z,r)}$ has the uniform distribution on the set of strings $\xi^n(t, z + 4, r)$. But the uniform distribution is the distribution of X conditional on $(T, Z, R) = (t, z + 4, r)$. That is, we have proven that

$$\mathcal{L}(\tilde{X}_{(t,z,r)}) = \mathcal{L}(X | (T, Z, R) = (t, z + 4, r)),$$

which finishes this proof. \blacksquare

Note that we have seen what happens with the variables T, Z, R after our random modification, let us see what happens with the length of the LCS after our random modification. In what follows, we always consider a triplet of values (t, z, r) such that $P((T, Z, R) = (t, z, r)) \neq 0$. For any $\epsilon > 0$ let $U_{t,r}^n(\epsilon)$ denote the event that the map

$$D_Z \rightarrow \mathbb{N} : z \mapsto L_n(t, z, r)$$

is increasing with a slope of at least $\epsilon/8$ on a scale $c_2 \ln(n)$ where $c_2 > 0$ is a large constant not depending on n . More precisely, $U_{t,r}^n(\epsilon)$ is the event that for any z_1, z_2 in D_Z , with $z_2 - z_1 \geq c_2 \ln(n)$ we have

$$L_n(t, z_2, r) - L_n(t, z_1, r) \geq (z_2 - z_1)\epsilon/8.$$

The event $U_{t,r}^n(\epsilon)$ has large probability because we assumed that inequality 2.1 holds. Hence $z \mapsto L_n(t, z, r)$ can be viewed somehow as behaving like a random walk with drift ϵ . In the next lemma we will show this looking at the event $U^n(\epsilon)$:

$$U^n(\epsilon) := \bigcap_{t \in D_T, r < l+1} U_{t,r}^n(\epsilon).$$

Lemma 2.6 *Given $\epsilon > 0$, take α from inequality 2.1 (theorem 1.2) and c_2 to be big enough but not depending on n , for example $c_2 \geq \frac{80}{\epsilon^2}$ depending on ϵ . Then, there exists a constant $k_* > 0$ not depending on n but on α and on c_2 such that:*

$$P(U^{nc}(\epsilon)) \leq \frac{k_*}{n^2} \quad (2.25)$$

for n large enough, provided 2.1 holds.

Proof. We are going to define an event $\mathcal{U}(\epsilon)$ for any $\epsilon > 0$. Let $\mathcal{U}_{(t,z,r)}(\epsilon)$ be the event that the expected conditional increase is larger than ϵ when we introduce the random change into $X_{(t,z,r)}$. More precisely, let $\mathcal{U}_{(t,z,r)}^n(\epsilon)$ be the event that

$$\mathbb{E}[L_n(t, z + 4, r) - L_n(t, z, r) | X_{(t,z,r)}, Y] \geq \epsilon \quad (2.26)$$

Let

$$\mathcal{U}^n(\epsilon) := \bigcap_{(t,z) \in D, r < l+1} \mathcal{U}_{(t,z,r)}^n(\epsilon).$$

hence

$$\mathbb{P}(\mathcal{U}^{nc}(\epsilon)) \leq \sum_{(t,z) \in D, r < l+1} \mathbb{P}(\mathcal{U}_{(t,z,r)}^{nc}(\epsilon)). \quad (2.27)$$

Note that inequality 2.1 provides a bound for the probability that the conditional expected increase of LCS due to our random modification not being larger or equal to ϵ . That probability bound is $\exp(-n^\alpha)$. The only problem is that the bound is for X and Y whilst the event $\mathcal{U}_{(t,z,r)}^n(\epsilon)$ is for $X_{(t,z,r)}$ and Y . By going on to conditional probability we must multiply the probability by $\mathbb{P}((T, Z, R) = (t, z, r))$. Hence we find

$$\mathbb{P}(\mathcal{U}_{(t,z,r)}^{nc}(\epsilon)) \leq \frac{\exp(-n^\alpha)}{\mathbb{P}((T, Z, R) = (t, z, r))}. \quad (2.28)$$

We can next use the lower bound on $\mathbb{P}((T, Z, R) = (t, z, r))$ provided by lemma 2.4 for all values $(t, z) \in D$ and $r < l + 1$ to inequality 2.28 and obtain

$$\mathbb{P}(\mathcal{U}_{(t,z,r)}^{nc}(\epsilon)) \leq \frac{1}{k_0} \cdot n \cdot \exp(-n^\alpha). \quad (2.29)$$

which still gives an exponentially small bound in n . Applying now 2.29 to inequality 2.27, we obtain

$$\mathbb{P}(\mathcal{U}^{nc}(\epsilon)) \leq \frac{4lc^2}{k_0} \cdot n^2 \cdot \exp(-n^\alpha). \quad (2.30)$$

Which is an exponentially small bound in n . Note that when the event $\mathcal{U}^n(\epsilon)$ holds, we have that $z \mapsto L_n(t, z, r)$ behaves like a random walk with drift ϵ . Let us formalize this. As before, let $\{z_0, z_1, z_2, \dots, z_d\}$ be the set for the admissible values of Z . For fixed $t \in D_T$ and $r < l + 1$, we are going to define $L_n^*(t, z)$ inductively for $z \in \{z_0, z_1, z_2, \dots, z_d\}$. Let us define $L_n^*(t, z, r) := L_n(t, z, r)$ for every $z \in \{z_0, z_1, z_2, \dots, z_d\}$. Given $\tilde{z} \in \{z_0, z_1, z_2, \dots, z_d - 4\}$ let us define $L_n^*(t, \tilde{z} + 4, r)$ as follows:

$$L_n^*(t, \tilde{z} + 4, r) = \begin{cases} L_n(t, \tilde{z} + 4, r) & \text{if } \mathcal{U}_{(t,s,r)}^n(\epsilon) \text{ hold for all } s \in \{z_0, z_1, \dots, \tilde{z}\} \\ L_n^*(t, \tilde{z}, r) + \epsilon & \text{otherwise} \end{cases}$$

Note that when the event $\mathcal{U}^n(\epsilon)$ holds, then $L_n(t, z, r)$ and $L_n^*(t, z, r)$ are identical for all $t \in D_T$, $r < l + 1$ and $z \in \{z_0, z_1, \dots, z_d\}$. Let $\mathcal{V}_{t,r}^n(\epsilon)$ be the event that the map

$$D_Z \rightarrow \mathbb{N} : z \mapsto L_n^*(t, z, r)$$

is increasing with a slope of at least $\epsilon/8$ on a scale $c_2 \ln n$.

Let $\mathcal{V}^n(\epsilon)$ be the event

$$\mathcal{V}^n(\epsilon) := \bigcap_{t \in D_T, r < l+1} \mathcal{V}_{t,r}^n(\epsilon).$$

Hence by using proposition 2.2 we have that:

$$\mathbb{P}(\mathcal{V}^{nc}(\epsilon)) \leq \sum_{t \in D_T, r < l+1} \mathbb{P}(\mathcal{V}_t^{nc}(\epsilon)) \leq \sum_{t \in D_T, r < l+1} 2n^{-\tau} \leq 4lc n^{0.5-\tau} \quad (2.31)$$

where $\tau = \frac{\epsilon^2 c_2}{32}$. When $\mathcal{U}^n(\epsilon)$ holds then $\mathcal{V}^n(\epsilon)$ and $U^n(\epsilon)$ are equivalent. Hence

$$\mathcal{U}^n(\epsilon) \cap \mathcal{V}^n(\epsilon) \subset U^n(\epsilon)$$

Hence by using 2.30 and 2.31 we get:

$$\mathbb{P}(U^{nc}(\epsilon)) \leq \mathbb{P}(\mathcal{U}^{nc}(\epsilon)) + \mathbb{P}(\mathcal{V}^{nc}(\epsilon)) \leq \frac{4lc^2}{k_0} \cdot n^2 \cdot \exp(-n^\alpha) + 4lc n^{0.5-\tau} \quad (2.32)$$

To show that the last inequality gives us a rate of convergence to zero as a constant divided by a polynomial in n , we try now to get a closed form for the inequality supposing extra information for the involved constants.

Taking $c_2 \geq \frac{80}{\epsilon^2}$ we have the following bound for the exponent:

$$0.5 - \tau \leq -2$$

therefore we can bound

$$4lc n^{0.5-\tau} \leq \frac{4lc}{n^2}. \quad (2.33)$$

Also, we have that:

$$n^2 \exp(n^{-\alpha}) \leq \frac{1}{n^2} \quad (2.34)$$

holds for n large enough. So, by using 2.33 and 2.34 in 2.32 we can finally bound:

$$\begin{aligned} \mathbb{P}(U^{nc}(\epsilon)) \leq \mathbb{P}(\mathcal{U}^{nc}(\epsilon)) + \mathbb{P}(\mathcal{V}^{nc}(\epsilon)) &\leq 4lc^2 \tilde{c}_2 n^2 \cdot \exp(-n^\alpha) + 4lc n^{0.5-\tau} \\ &\leq (4lc^2 \tilde{c}_2 + 4lc) \cdot \frac{1}{n^2} \end{aligned}$$

for n large enough, which ends the proof with $k_* = 4lc^2 k_0 + 4lc$. \blacksquare

Proposition 2.2 *Given $\epsilon > 0$, let $\mathcal{V}_{t,r}^n(\epsilon)$ denote the event that the map $z \mapsto L^*(t, z, r)$ is increasing with a slope at least $\epsilon/8$ on a scale $c_2 \ln(n)$. Given $t \in D_T$, $r < l+1$ and $z_1, z_2 \in D_Z$ such that $z_2 - z_1 \geq c_2 \ln(n)$ we have the following inequality:*

$$\mathbb{P}(\mathcal{V}_{t,r}^{nc}(\epsilon)) \leq 2n^{-\tau}$$

where $\tau = \frac{\epsilon^2 c_2}{32}$.

Proof. Let $z_1, z_2 \in D_Z$ such that $z_1 < z_2$. In order to simplify the notation, let us assume that $z_2 - z_1$ can be divided by 4 and denote $\frac{z_2 - z_1}{4} = m \in \mathbb{N}$. Let z_0 be the most left point of D_Z . Given $\epsilon > 0$, let us remember that $\mathcal{V}_{t,r}^n(\epsilon)$ is the event such that the following inequality holds:

$$L^*(t, z_2, r) - L^*(t, z_1, r) \geq \frac{\epsilon}{8}.$$

Now let us define the filtration $\mathfrak{F}_0 \subset \mathfrak{F}_1 \subset \dots \subset \mathfrak{F}_m$ as follows:

$$\mathfrak{F}_i := \sigma \left(X_{(t, z_0, r)}, X_{(t, z_1, r)}, \dots, X_{(t, z_1 + 4i, r)}; Y \right)$$

for $i = 1, \dots, m$. Let us denote

$$e_i = E[L_n^*(t, z_1 + 4(i+1), r) - L_n^*(t, z_1 + 4i, r) \mid \mathfrak{F}_i]$$

and define a martingale M_0, M_1, \dots, M_m with respect to the filtration $\mathfrak{F}_0 \subset \mathfrak{F}_1 \subset \dots \subset \mathfrak{F}_m$ as follows:

$$\begin{aligned} M_0 &:= L_n^*(t, z_1, r) \\ M_{i+1} - M_i &:= L_n^*(t, z_1 + 4(i+1), r) - L_n^*(t, z_1 + 4i, r) - e_i \end{aligned}$$

for $i = 1, \dots, m$. By definition of the map $z \mapsto L_n^*(t, z, r)$ we have an expected increase of at least ϵ every time z gets increased by 4, so that the expected increase of

$$E[L_n^*(t, z_1 + 4(i+1), r) - L_n^*(t, z_1 + 4i, r)]$$

is at least ϵ which implies that the following inequality

$$e_i \geq \epsilon \tag{2.35}$$

is satisfied almost surely for every $i = 1, \dots, m$. We can write the increase of the map $z \mapsto L_n^*(t, z)$ in terms of the martingale M_0, \dots, M_m in the following way:

$$L^*(t, z_2, r) - L^*(t, z_1, r) = M_m - M_0 + \sum_{i=0}^{m-1} e_i \tag{2.36}$$

Now, we are ready to estimate the probability of $\mathcal{V}_{t,r}^{nc}(\epsilon)$:

$$\begin{aligned} \mathbb{P}(\mathcal{V}_{t,r}^{nc}(\epsilon)) &= \mathbb{P}\left(L^*(t, z_2, r) - L^*(t, z_1, r) \leq \frac{\epsilon}{8}(z_2 - z_1)\right) \\ \text{(by equality 2.36)} &\leq \mathbb{P}\left(M_m - M_0 + \sum_{i=0}^{m-1} e_i \leq \frac{\epsilon}{8}(z_2 - z_1)\right) \\ &= \mathbb{P}\left(M_m - M_0 \leq \frac{\epsilon}{8}(z_2 - z_1) - \sum_{i=0}^{m-1} e_i\right) \\ \text{(by 2.35 and } z_2 - z_1 = 4m) &\leq \mathbb{P}\left(M_m - M_0 \leq \frac{\epsilon}{8}(z_2 - z_1) - \frac{\epsilon}{4}(z_2 - z_1)\right) \\ &= \mathbb{P}\left(M_m - M_0 \leq -\frac{\epsilon}{8}(z_2 - z_1)\right) \end{aligned} \tag{2.37}$$

At this point we want to use Azuma-Hoeffding inequality 1.3. For this, we note that for every $i = 1, \dots, m$ we have

$$\mathbb{P}(|M_{i+1} - M_i| \leq 1) = 1$$

since $\epsilon < 1$ and we take $v = \frac{\epsilon}{8}(z_2 - z_1)$ for writing down:

$$\begin{aligned} \mathbb{P}\left(M_m - M_0 \leq -\frac{\epsilon}{8}(z_2 - z_1)\right) &\leq 2 \exp\left(-\frac{v^2}{2m}\right) \\ (\text{by using } z_2 - z_1 = 4m) &= 2 \exp\left(-\frac{\epsilon^2}{32}(z_2 - z_1)\right) \end{aligned} \quad (2.38)$$

Combining together 2.37 and 2.38 we finally have:

$$\mathbb{P}(\mathcal{V}_{t,r}^{nc}(\epsilon)) \leq 2 \exp\left(-\frac{\epsilon^2}{32}(z_2 - z_1)\right)$$

from where, after taking $z_2 - z_1 \leq c_2 \ln(n)$, we have:

$$\mathbb{P}(\mathcal{V}_{t,r}^{nc}(\epsilon)) \leq 2 \exp\left(-\frac{\epsilon^2 c_2}{32} \ln(n)\right) = 2 n^{-\frac{\epsilon^2 c_2}{32}}$$

which finishes the proof \blacksquare

Note that by law of total probability $\mathbb{E}[\text{VAR}[L_n(T_D, Z_D, R)|T_D, R]]$ is equal to :

$$\mathbb{P}(U^n(\epsilon))\mathbb{E}[\text{VAR}[L_n(T_D, Z_D, R)|T_D, R] | U^n(\epsilon)] + \mathbb{P}(U^{nc}(\epsilon))\mathbb{E}[\text{VAR}[L_n(T_D, Z_D, R)|T_D, R] | U^{nc}(\epsilon)],$$

for every $\epsilon > 0$ and hence:

$$\mathbb{E}[\text{VAR}[L_n(T_D, Z_D, R)|T_D, R]] \geq \mathbb{P}(U^n(\epsilon))\mathbb{E}[\text{VAR}[L_n(T_D, Z_D, R)|T_D, R] | U^n(\epsilon)] \quad (2.39)$$

Now, conditional on the event $U^n(\epsilon)$ holding, we have that the random map:

$$D_Z \rightarrow \mathbb{N} : z \mapsto L_n(t, z, r)$$

has a slope of at least $\epsilon/8$ on a scale of $c_2 \ln(n)$ (as in proposition 2.2) for any $t \in D_T$ and $r < l + 1$, then:

$$\begin{aligned} z_2 - z_1 \geq c_2 \ln(n) &\Rightarrow L_n(t, z_2, r) - L_n(t, z_1, r) \leq \frac{\epsilon}{8}(z_2 - z_1) \\ z_2 - z_1 < c_2 \ln(n) &\Rightarrow L_n(t, z_2, r) - L_n(t, z_1, r) \leq 2(z_2 - z_1) \end{aligned}$$

hold. Hence, conditional on $U^n(\epsilon)$, we can apply lemma 2.2 and obtain:

$$\text{VAR}[L_n(t, Z_D, R)|T_D = t, R = r, U^n(\epsilon)] \geq \frac{\epsilon^2}{64} \left(1 - 16 \frac{(\epsilon/8 + 2)c_2 \ln(n)}{\epsilon \sqrt{\text{VAR}[Z_D|T_D = t, R = r]}}\right) \text{VAR}[Z_D|T_D = t, R = r] \quad (2.40)$$

The next results give us an uniform bound for $\text{VAR}[Z_D|T_D = t, R = r]$ for all $t \in D_T$.

Lemma 2.7 *There exists a constant $K > 0$ not depending on n such that:*

$$1 - \frac{K}{\sqrt{n}} \leq \frac{P(Z_D = z + 4 | T_D = t, R = r)}{P(Z_D = z | T_D = t, R = r)} \leq 1 + \frac{K}{\sqrt{n}} \quad (2.41)$$

for every $(t, z) \in D$, $r < l + 1$ and n large enough.

Lemma 2.8 *There exists a constant $C > 0$ not depending on n such that:*

$$\text{VAR}[Z_D | T_D = t, R = r] \geq C \cdot n$$

for every $t \in D_T$, $r < l + 1$ and for every n large enough.

Using the bound in lemma 2.8 we get the following inequality:

$$\left(1 - 16 \frac{(\epsilon/8 + 2)c_2 \ln(n)}{\epsilon \sqrt{\text{VAR}[Z_D | T_D = t, R = r]}} \right) \geq \left(1 - 16 \frac{(\epsilon/8 + 2)c_2}{\epsilon \sqrt{C}} \cdot \frac{\ln(n)}{\sqrt{n}} \right) \geq 0.5 \quad (2.42)$$

for n large enough. Using inequality 2.42 above with inequality 2.40 we find:

$$\text{VAR}[L_n(t, Z_D, R) | T_D = t, R = r, U^n(\epsilon)] \geq \frac{\epsilon^2}{64} 0.5 \cdot \text{VAR}[Z_D | T_D = t, R = r].$$

Using again lemma 2.8 we find that the left side of the above inequality is larger than $\frac{C\epsilon^2}{128}n$ and hence:

$$\mathbb{E}[\text{VAR}[L_n(T_D, Z_D, R) | T_D, R] | U^n(\epsilon)] \geq \frac{C\epsilon^2}{128}n \quad (2.43)$$

We can now combine inequalities 2.19, 2.20, 2.21, 2.39 and 2.43 to obtain:

$$\text{VAR}[L_n] \geq P(U^n(\epsilon)) \frac{C\epsilon^2}{1000}n$$

and plugging in the lower bound for $P(U^n(\epsilon))$ obtained in 2.25 (lemma 2.6) we get:

$$\text{VAR}[L_n] \geq \frac{C\epsilon^2}{1000}n \left(1 - \frac{k_*}{n^2} \right)$$

with $k_* > 0$ is the constant from lemma 2.6. This expression is a lower bound of order $\Theta(n)$ for $\text{VAR}[L_n]$. Hence, we have finished proving the statement of the result in theorem 1.2.

Acknowledgments

The authors would like to thank the support of the German Science Foundation (DFG) through the International Graduate College "Stochastics and Real World Models" (IRTG 1132) at Bielefeld University and through the Collaborative Research Center 701 "Spectral Structures and Topological Methods in Mathematics" (CRC 701) at Bielefeld University.

References

- [1] V. Chvatal and D. Sankoff. *Longest common subsequences of two random sequences*. J. Appl. Probability, 12 : 306–315, 1975.
- [2] M. S. Waterman. *Estimating statistical significance of sequence alignments*. Phil. Trans. R. Soc. Lond. B, 344:383-390, 1994.
- [3] H. Matzinger, Torres, F. *Fluctuation of the longest common subsequence for sequences of independent blocks*. Submitted, 2010.
- [4] Torres, F. *On the probabilistic longest common subsequence problem for sequences of independent blocks*. Ph.D. thesis, University of Bielefeld. March 2009. Online <http://bieson.ub.uni-bielefeld.de/volltexte/2009/1473/>
- [5] J. Lember, H. Matzinger. *Standard Deviation of the Longest Common Subsequence*. Ann. Probab. Volume 37, Number 3: 1192-1235, 2009.
- [6] F. Bonetto and H. Matzinger. *Fluctuations of the longest common subsequence in the case of 2- and 3-letter alphabets*. Latin American Journal of Probability and Mathematics, Volume 2:195–216, 2006.
- [7] G. Grimmett. and D. Strizaker. *Probability and Random Processes*, Oxford University Press, 2001. Third edition.
- [8] S.M. Ross, *Introduction Probability Models*. Academic Press, 8 edition, 2002.